This manuscript was published as Liesefeld, H.R., Liesefeld, A.M., & Müller, H.J. (2019). Two good reasons to say 'change!' – ensemble representations as well as item representations impact standard measures of VWM capacity. *British Journal of Psychology, 110*, 328-356. doi:10.1111/bjop.12359

Two good reasons to say 'change!' – ensemble representations as well as item representations impact standard measures of VWM capacity

Heinrich René Liesefeld[1], Anna M. Liesefeld[1], Hermann J. Müller[1,3]

[1]Department Psychologie, Ludwig-Maximilians-Universität, München, Germany; [2]Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Germany; [3]Department of Psychological Sciences, Birkbeck College, University of London, UK

Author Note

This work was supported by the German Research Foundation (DFG) under Grant MU773/14-1, awarded to HJM, by LMU Munich's Institutional Strategy LMUexcellent within the framework of the German Excellence Initiative, and by the Graduate School of Systemic Neurosciences, Munich Center for Neurosciences – Brain & Mind.

Correspondence concerning this article should be addressed to Heinrich René Liesefeld Department Psychologie, Ludwig-Maximilians-Universität, Leopoldstr. 13, D-80802 Munich, Germany, E-mail: Heinrich.Liesefeld@psy.lmu.de

Abstract

Visual working memory (VWM) is a central bottleneck in human information processing. Its capacity is most often measured in terms of how many individual-item representations VWM can hold (*k*). In the standard task employed to estimate *k*, an array of highly discriminable colour patches is maintained and, after a short retention interval, compared to a test display (change detection). Recent research has shown that with more complex, structured displays, change-detection performance is, in addition to individual item representations, supported by ensemble representations formed as a result of spatial sub-groupings. Here, by asking participants to additionally localise the change, we reveal indication for an influence of ensemble representations even in the very simple, unstructured, displays of the colour-patch change-detection task. Critically, pure-item models from which standard formulae of *k* are derived do not consider ensemble representations and, therefore, potentially overestimate *k*. To gauge this overestimation, we develop an item-plus-ensemble model of change detection and change localisation. Estimates of *k* from this new model are about 1 item (~30%) lower than the estimates from traditional pure-item models, even if derived from the same data sets.

Keywords. visual short-term memory, working memory capacity, change detection, change localisation, cognitive modelling

Two good reasons to say 'change!' – ensemble representations as well as item representations impact standard measures of VWM capacity

The most characteristic feature of working memory (WM) is its limited capacity. This limit delineates it from sensory memory as well as long-term memory. Furthermore, people considerably differ in their WM capacities and their individual limits are predictive of all kinds of highly relevant everyday skills and developmental trajectories (Alloway & Alloway, 2010; Conway, Kane, & Engle, 2003; Fukuda, Vogel, Mayr, & Awh, 2010; Luck & Vogel, 2013). No wonder, therefore, that the characterization and measurement of the WM capacity limit has attracted a huge research effort. This effort was most sophisticated and debates on the nature of WM representations most vigorous in the realm of *visual* working memory (*VWM*). One reason why investigating VWM holds particular promise for gaining insights into the nature of WM capacity limits is that VWM tasks, such as *change detection*, afford tight control over stimulus characteristics, task strategies and other potentially confounding factors (Cowan, 2001; Luck & Vogel, 1997).

In the most common version of the change-detection task, pioneered by Pashler (1988) and popularized by Luck and Vogel (1997), a memory array containing several colour patches is presented for a few hundred milliseconds, followed by a retention interval of about one second, and then the presentation of a test array until response. On each trial, the test array is either identical to the memory array or differs in one colour, and the task is to indicate whether or not something has changed. This has become a standard task to measure VWM capacity in terms of the number of item representations VWM can maximally hold, and estimates of VWM capacity derived from this task typically range from $k = 3$ to $k = 4$ representations on average (Alvarez & Cavanagh, 2004; Awh, Barton, & Vogel, 2007; Brady & Alvarez, 2015b; Cowan, 2001; Cowan et al., 2005; Fukuda et al., 2010; Luck, 2008; Luck & Vogel, 2013; Vogel & Machizawa, 2004).

There is disagreement on what exactly an item representation in VWM comprises: whole objects versus single features (Luck & Vogel, 1997; Fougnie & Alvarez, 2011; Fougnie, Cormica, & Alvarez,

2013; Wheeler & Treisman, 2002), and whether VWM capacity is best described as a fixed number of item slots or a continuous resource (Bays, 2014, 2015; Bays & Husain, 2008; Ma, Husain, & Bays, 2014; van den Berg, Awh, & Ma, 2014; van den Berg, Shin, Chou, George & Ma, 2012). However, across all these camps, researchers assume that VWM holds representations of individual items and that the colour patches in the above-described change-detection task are simple enough to be contained in one representation each; that is, according to all current theories, participants in the standard, simple change-detection task maintain (some) colour patches as separate entities in VWM.

But what if additional representations contribute to performance on the change-detection task, thus confounding capacity estimates? Brady and colleagues (Brady & Alvarez, 2015b; Brady & Tenenbaum, 2013) found that, in certain situations, estimates of VWM capacity can be heavily biased upwards, because participants detect some changes based on *ensemble representations* of the scene even if the individual changed item was not represented separately in VWM. "Ensemble representation" is an umbrella term for all those representations that do not conform to the classical idea of separate entities in VWM (e.g., individual colours). These representations contain information on what several (not necessarily all) memorized objects have in common, such as their average feature value (Alvarez, 2011; Alvarez & Brady, 2015a; Alvarez & Oliva, 2009; Brady, Konkle, & Alvarez, 2011; Brady & Tenenbaum, 2011; Chong, Joo, Emmanouil, & Treisman, 2008; Chong & Treisman, 2003, 2005). If such ensemble representations support performance in even the simplest version of the change-detection task described above, pure-item models (which are implicitly accepted by calculating *k*) would routinely overestimate VWM capacity, and an item-plus-ensemble model would be needed to obtain an unbiased estimate of the number of individual item representations stored in VWM.

Brady and Alvarez (2015b) devised an elegant way to demonstrate the influence of ensemble representations by experimentally manipulating their usefulness to solve a given task. Replicating earlier work (Awh et al., 2007), they had participants memorize a mixture of Chinese characters and 3-dimensional cubes and introduced two types of changes (see Fig. 1A): within-category changes (e.g., cube → cube) or between category changes (e.g., cube → character). Importantly, capacity estimates turned out

much lower for within- than for between-category changes. They explained this finding in terms of spatial ensemble representations (textures) formed based on some type of grouping (segmentation) of the display: several objects belonging to the same category may be represented by a single ensemble representation, for example, when several Chinese characters form a cluster in the upper left quadrant of the memory display; accordingly, when a cube is shown at one of these cluster positions at test, the change may not be detected based on a representation of the individual character that was replaced by the cube, but rather based on 'knowledge' that all items in this display region were characters. Furthermore, in the whole-display change-detection task introduced above, people might often not notice a change in any specific individual item (e.g., from Chinese character to cube), but a change in the overall spatial composition (e.g., in terms of overall luminance within a given region). To demonstrate the involvement of ensemble representations, Brady and Alvarez created more heterogeneous memory displays by drawing objects from four (instead of just two) different categories, thereby reducing opportunities for grouping. As a result, and as predicted by the hypothesis of spatial-ensemble representations, capacity estimates for between-category changes exhibited a marked decrease. Such findings advise caution with regard to VWM capacity estimates derived from non-standard displays (often featuring complex objects, as in Fig. 1A) that can be structured into sub-groups (as carried to the extreme in Fig. 1B) and thus provide ready opportunities for employing spatial ensemble representations (Brady & Tenenbaum, 2013).

A simple lesson one might draw from these findings is that employing such *non-standard* versions of the change-detection task should be avoided if the goal is to obtain a valid measure of VWM capacity. However, even the most *standard* measure of VWM capacity (i.e., *k*, which was introduced above), derived from the most *standard* change-detection task (i.e., the colour-patch version introduced above), might be contaminated by ensemble representations. Potentially, colour repetitions in the memory display allow for similar grouping strategies as in Brady and colleagues' work (see Figure 1C; Brady & Tenenbaum, 2013). A simple solution to avoid such biasing of VWM capacity estimates would be to avoid colour repetitions within displays (Fig. 1D). Still, even without repetitions (and thus obvious opportunities for sub-grouping), other ensemble representations, such as the 'average' colour or luminance

of the whole display or (less obvious) sub-groups of objects, might be at work in the colour-patch change-detection task, thus potentially contaminating estimates of VWM capacity. Without knowing exactly what these representations look like (how is colour 'averaged'?; based on what criterion are sub-groups formed?), Brady and colleagues' elegant approach of manipulating the characteristics of the memory display accordingly (Brady & Alvarez, 2015b; Brady & Tenenbaum, 2013; see also Brady & Alvarez, 2011, 2015a; Brady, Konkle, & Alvarez, 2009) has its limitations and an alternative approach is needed.

The approach evaluated here aims to explicitly measure the incidence of the hypothesized two forms of representations, that is, the incidence of item-based and ensemble-based change detections, respectively. The logic underlying this approach is as follows: whenever observers detect that a specific item has changed, they should be able to select the changed item from among all items in the test display, that is, to localise the change (making a point-and-click response with a computer mouse; for elaborations on this assumption, see the General Discussion).[1] By contrast, whenever detection is based on an ensemble representation of a group of items – so that change information would be 'lumped across' several objects, each of which could, in principle, have changed – observers would have to resort to guessing which specific item has changed.

Given this, involvement of ensemble representations in the classical colour-patch change-detection task would predict that there is a considerable incidence of trials on which participants detect a change but are unable to localise it, whereas – excluding detection responses that simply happen to be correct as a result of random guessing – such trials should not occur when only individual-item representations are involved. In fact, incidences of change detection or change classification without change localisation are commonly observed in the change-blindness literature (e.g., Agostinelli et al., 1986; Ball & Busch, 2015; Busch, Dürschmid, & Herrmann, 2010; Becker, Pashler, & Anstis, 2000; Turatto & Bridgeman, 2005; Hughes, Caplovitz, Loucks, & Fendrich, 2012). For example, people can correctly classify whether the

[1] This is, of course, not the first study to employ a 'change-localisation' task with simple colour patches (see, e.g., Gold et al., 2018; Johnson et al., 2013; Kornblith, Buschman, & Miller, 2016; Shin & Ma, 2017; van den Berg et al., 2012). And, in fact, Pailian and Halberda (2015, Exp. 3) combined a standard colour-change-detection task with a localisation task comparable to our Experiment 1. However, the present study is – to our knowledge – the first to use change localisation to estimate the incidence of ensemble-based change detections.

average emotion in a set of faces changed towards happier or angrier without being able to select a changed face (Haberman & Whitney, 2011). A well-known phenomenon in this literature is that people often report 'sensing' a change before they are able to report where the change is (Rensink, 2004). Others refer to this phenomenon as detecting a change in the 'gist' of the scene (Haberman & Whitney, 2011). 'Sensing' a change and detecting a change in the 'gist' of a scene might just be detection of a change based on ensemble representations (Ball & Busch, 2015; Haberman & Whitney, 2011). However, change blindness studies differ in many respects from the typical colour-change detection task: they typically involve a higher incidence of change trials (often up to 100%, as a result of which change-detection accuracy cannot be measured) and shorter retention intervals (implying the involvement of iconic memory), as well as using more complex stimuli (thus providing more opportunities for using ensemble representations by grouping based on sub-categories or other principles).

Thus, the main issue addressed here is whether ensemble representations are involved in a standard colour-patch change-detection task even without colour repetitions. As we did obtain evidence for the involvement of ensemble representations, we further (a) explored the collected data for information on the nature of these ensemble representations and (b) developed a model that assumes both item and ensemble representations and takes both change-detection and -localisation performance into account. With regard to modelling, we compare estimates of VWM capacity in terms of the number of individually stored items (*k* index) from our proposed (item-plus-ensemble-representation-based) model to those from a pure-item model and find that not taking ensemble representations into account (i.e., the latter model) yields a substantial overestimation of the number of individual items stored in VWM.

Experiment 1

Methods

Participants. Students recruited at our university participated in this study (n = 18, 1 left-handed, median age: 20.5 years, range: 18-30 years, 14 female). All participants had normal or corrected to normal vision. They gave prior informed consent (in writing) and received course credit or were paid for their participation.

Stimuli, design, and procedure. Participants performed a colour-change-detection task, as illustrated in Figure 2. Stimuli were 9 easily discriminable coloured squares (red, green, blue, yellow, pink, turquoise, black, white, and orange), 0.75° of visual angle in size, which were presented against a dark grey background on a CRT monitor (screen resolution: 1,024 × 768 pixels, refresh rate: 120 Hz).

The memory array consisted of six coloured squares and was presented for 200 ms. Each colour appeared only once per trial and squares were placed at random positions inside a (virtual) rectangular frame, 10° (width) × 9° (height) in size (and centred on fixation), with the restriction that adjacent squares were separated (centre-to-centre) by at least 1.5° of visual angle. Following the memory array, there was a 900-ms retention interval, and then a test array appeared. Participants pressed a mouse key (using their left or right index finger; counterbalanced) to indicate whether or not a colour had changed between the memory and the test array; they were told to produce this response as accurately as possible. In half the trials (order randomized), one of the objects changed its colour into one of the colours that had not appeared in the memory array; in the other half, all colours remained the same. A white fixation cross (0.5° in size) was present throughout. After a response was registered, the fixation cross turned green (correct answer) or red (incorrect answer) for 1,000 ms. When no response was given within 2,500 ms, the trial was aborted and the fixation cross turned blue (time-out).

On change trials, after having made their change-detection response (i.e., regardless of whether the response was “change” or “no change”), participants performed the additional task of indicating which object they believed had changed. They did this by moving the mouse cursor onto one of the objects in the test display and clicking one of the mouse keys, without time pressure. The fixation cross then again turned green (correct response) or red (incorrect response) for 1,000 ms. The inter-trial interval was jittered between 800 and 1,600 ms.

Participants first performed a block of 20 practice trials (without response deadline), followed by 9 blocks of 50 regular trials each – making a total of 225 change and 225 no-change trials. One data set was incomplete due to technical problems (only 150 instead of 450 regular trials)[2].

Data analysis. One difficulty with estimating the incidence of correct localisations when the change is detected is that the true detection rate is not directly measurable. A "change" response on a change trial (which is referred to as a 'hit', in contrast to a 'miss') can occur for two reasons: the observer did detect the change or s/he just guessed correctly[3]. Fortunately, formulae are available to estimate the true detection rate ($\hat{d}$) from the observed behaviour by taking into account trials on which no change was present and the observer responded "change" erroneously ('false alarms'). For the whole-report change-detection task employed here, with $\hat{h}$ = hit rate and $\hat{f}$ = false-alarm rate, this formula is:

$$\hat{d} = \frac{\hat{h} - \hat{f}}{1 - \hat{f}} \quad (1)$$

Although Pashler (1988) and Rouder, Morey, Morey, and Cowan (2011) interpret $\hat{d}$ as the probability that the changed object was in VWM (and the change was detected as a consequence), a simpler, less model-specific interpretation is that $\hat{d}$ directly reflects the true detection rate, also containing detections based on ensemble representations. If, in the latter case, the false-alarm rate ($\hat{f}$) reflects some base rate of guessing "change" that also applies when a change was present but not detected, this results in the same formula (Eq. 1, see Appendix A). Based on this formula, we can estimate the percentage of hits that are due to true detections (instead of guessing) and compare localisation performance ($\hat{l}$) on hit trials against this percentage.

We use a relatively large set size of 6 items (in the memory and test displays), which is considerably above typical VWM capacity estimates (3-4 items; Cowan, 2001; Luck & Vogel, 2013). This has two main advantages: First, it becomes easier to differentiate between simple guessing on the change-

[2] Excluding this data set had no noteworthy effect on the results.

[3] Alternatively, the observer might 'detect' a change due to a failure of VWM: If the wrong colour is remembered at a certain position, the (non-changed) object at that position in the test display might appear to have changed (see the swapping model described below). Such apparent change detections would occur equally often on 'change' and 'no change' trials and, thus, be treated as guessing in the formula below.

localisation part and non-perfect but informed performance (any localisation performance above 1/6 = 17% correct is above chance with set size 6, whereas with, e.g., set size 2, only performance above 50% would be above chance). Second, in case participants had difficulty selecting the correct location, we can examine for possible biases in selecting incorrect objects (e.g., to examine whether incorrect objects close to the target location were selected disproportionately more often).

Which object is erroneously selected might be diagnostic with regard to the type of ensemble representations employed: detecting a change in the spatial ensemble (texture) representations examined by Brady and Alvarez (2015b) would provide at least approximate information about the location of the change; participants would know which region of the display has changed its texture and select an item within this region instead of picking one further away from the change (e.g., 'somewhere in the upper right corner'). Consequently, incorrect localisation responses would be closer to the change than expected by chance. If erroneous localisations would turn out independent of the location of the change, change-detection was probably based on non-spatial ensemble representations of the whole display (e.g., the 'average' colour across all objects; Brady & Alvarez, 2011, 2015a; Brady et al., 2011). Our data provide evidence for the first alternative (spatial ensemble representations).

All reported *t* tests were two-tailed and, additionally, the respective effect sizes (d_z and *d*; Cohen, 1988) and Bayes Factors (BF; Rouder, Speckman, Sun, Morey, & Iverson, 2009) are given. BFs (objective Jeffrey-Zellner-Siow Prior with a scale parameter of 0.707) provide information on the evidence for the null hypothesis of equivalence (BF_{01}) or the alternative hypothesis of difference (BF_{10}) of the two values compared; BFs > 3 indicate substantial evidence.

Results

Change-detection performance. Participants responded correctly on 61% of all change trials (hit rate, $\hat{h} = .61$). They pressed "change" when there was none in only 10% of trials (false-alarm rate, $\hat{f} = .10$), yielding an estimated detection rate of 56% ($\hat{d} = .56$, Eq. 1). That is, participants used a reasonably conservative criterion to indicate that a change had occurred (i.e., they guessed rarely) and truly detected the change on more than half the trials. This results in an average capacity estimate of $\hat{k}_p$ =

3.38 items according to the standard pure-item model (Pashler, 1988; Rouder et al., 2011), which lies within the usual range (3 to 4 items; Luck & Vogel, 1997, 2013). In sum, we observed the performance pattern that would be expected in this type of change-detection task.

Localisation performance and hypothesis tests. Pure-item models would predict that participants are able to select the changed object on all detection trials (see General Discussion for rebuttals of various alternative explanations). The percentage of correct responses on change trials that are due to true detections of the change (rather than correct guesses) is 91% on average ($\frac{\hat{d}}{\hat{h}}$ = .91). Thus, if participants knew which object had changed whenever they truly detected a change, they should correctly select the changed object on at least $\frac{\hat{d}}{\hat{h}}$ of all hit trials (which is the lower bound and thus a conservative estimate for the present purposes, because this number would increase further when guessing on the localisation part was also taken into account). At variance with this, the correct object was indicated on only 71% of hit trials ($\hat{l}|hit$ = .71), which is significantly lower than the lower bound ($\frac{\hat{d}}{\hat{h}}$) predicted by a pure-item model (even when taking guessing on the change-detection part into account), $t(17) = 13.75$, $p < .001$, $d_z = 3.24$, $BF_{10} > 73 \times 10^6$. Interestingly, even when participants responded that they had not perceived a change, they selected the changed object at a rate of 31% ($\hat{l}|miss$ = .31), which is clearly above chance (1/6 = .17), $t(17) = 10.92$, $p < .001$, $d_z = 2.57$, $BF_{10} > 27 \times 10^5$, indicating that they knew that some of the items had (likely) not changed and chose among the remaining items (*informed guessing*; see Appendix A for an implication of this finding).

Model fitting. Another way to prove the insufficiency of a pure-item model to account for the observed localisation performance is to compare its quantitative predictions against the respective empirical data. In this approach, most underlying assumptions must be spelled out and the degree of misfit can be quantified, so that one can directly compare how well the data are explained by various competing models. The disadvantage is that our (current) state of understanding does not allow all model aspects to be specified, so that some assumptions must be made on a poor knowledge base. This also implies that any explicit model is only one possible instance from the respective family of models and future research

might find other instances that do a better job at explaining the data. By contrast, testing specific predictions derived from the assumption that only item representations contribute to change detection, as done above for Experiment 1 and below for the follow-on experiments, depends on relatively few specific assumptions, thus providing more general evidence against the whole family of pure-item models (including discrete-slot and continuous-resource versions). Accordingly, given their respective strengths and weaknesses, the two approaches complement each other.

Table 1 provides a summary of our model-fitting results; the models and the fitting procedure are detailed in Appendix B. In addition to one instantiation of a pure-item model and an item-plus-ensemble model, we also built a model that explains some of the gap between hit rate and localisation rate by swaps of colours between items in VWM (Bays, 2016; Bays, Wu, & Husain, 2009; Schneegans & Bays, 2019), instead of ensemble representations. The item-plus-ensemble-model fits the average performance data best (cumulative deviation: 0.03), with only a slight underestimation of the localisation rate on miss trials. One possible reason for this underestimation is that (in contrast to the model assumptions; see Appendix B) observers do make use of ensemble representations during guessing.

Next is the swapping model (cumulative deviation: 0.09), which overestimates the localisation rate on miss trials *and* underestimates the hit rate. Hit rate is likely underestimated because swapping also creates false alarms (when two items swap in memory on no-change trials, observers perceive two apparent changes) and the maximal swapping rate that is compatible with the observed false-alarm rate is insufficient to completely explain the full difference between hit rate and localisation rate. In other words, the swapping rate needed to produce the observed localisation rate and hit rate would produce a much higher false-alarm rate than is actually observed, namely $\hat{f} = .15$ instead of the observed $\hat{f} = .10$. An overestimation of k might (in part) explain the overestimation of localisation rate on miss trials, because too many items are excluded from (informed) guessing.

The pure-item model performs worst (cumulative deviation: 0.16). Most notably, it underestimates the hit rate and overestimates the localisation rate. As detailed above, pure-item models cannot concurrently explain a relatively high hit rate and a relatively low localisation rate, because both are

determined by the same *k* parameter. In the present case, the fitting algorithm settled on a compromise, so that both predictions considerably deviate from the empirically observed values.

Exploratory analyses. As detailed above, spatial ensemble/texture representations (Brady & Alvarez, 2015b; Brady & Tenenbaum, 2011) would likely provide observers with rough knowledge of the spatial region where the change occurred. To test whether our participants possessed such knowledge, we calculated the ordinal distances between the changed object and the erroneously selected object on trials with incorrect selections (i.e., whether the erroneously selected object was the closest, the second-closest, etc. to the changed object in the display). On hit trials with incorrect selections (when spatial ensemble representations were probably at play), this distance should be lower than the average ordinal distance between any two objects, which for a set size of $S = 6$ is $\frac{\sum_{i=1}^{S-1} i}{S-1} = 3$. Indeed, the observed average ordinal distance of the selected object on hit trials (2.65) was lower than this chance value, $t(17) = 5.10$, $p < .001$, $d = 1.20$, $BF_{10} > 313$. Furthermore, if some hits were based on spatial ensemble representations, incorrectly selected objects on hit trials should be closer to the changed object than incorrectly chosen objects on miss trials (on which neither item nor ensemble representations indicated a change [with a sufficient strength]), which was also the case, $t(17) = 3.15$, $p = .006$, $d_z = 0.74$, $BF_{10} = 8.35$. In fact, the distance on miss trials (2.87) did not differ significantly from chance, $t(17) = 1.99$, $p = .063$, $d = 0.47$, $BF_{10} = 1.21$. Figure 3 provides a more detailed picture of this spatial bias in erroneous selections.

The data from Experiment 1 also provide the interesting opportunity to examine whether representations of the change location smear out across time, so that localisations become more imprecise. In particular, when participants responded slowly on the change detection task, more time had passed since test-display onset before a localisation response could be issued. Splitting trials into those with fast and slow change-detection responses (median split) provided no evidence of such smearing across time, $t(17) = 0.71$, $p = .486$, $d_z = 0.17$, $BF_{01} = 3.29$. However, participants produced more hits, $t(17) = 2.11$, $p = .050$, $d_z = 0.50$, $BF_{10} = 1.44$, and more false alarms, $t(17) = 8.31$, $p < .001$, $d_z = 1.96$, $BF_{10} > 72 \times 10^3$, on slow trials, whereas detection rate did not significantly differ between fast and slow trials, $t(17) = 1.20$, $p = .248$, $d_z = 0.28$, $BF_{01} = 2.22$. This indicates an increased guess rate on slow trials. Furthermore, directly

estimating the incidence of ensemble-based detections (Eq. 8 in Appendix A) indicated that ensemble representations play a bigger role on trials with slow change-detection responses (21% vs. 9%), $t(17) = 3.87$, $p = .001$, $d_z = 0.91$, $BF_{10} = 31.25$.

Discussion

In a standard colour-change-detection task, we additionally asked participants to select the changed object from the memory array. Results revealed that participants often detect changes without being able to localise them. Importantly, localisation rate was lower than can be explained by guessing on the change-detection part. This provides support for the assumption that changes are often detected based on 'sensing' (Rensink, 2004) or ensemble statistics (Brady & Alvarez, 2011, 2015a,b; Brady & Tenenbaum, 2013). This interpretation was further corroborated by comparing various specific models: an item-plus-ensemble model provided the best account of the empirical data and was superior to a pure-item model as well as a model assuming that individual-item representations sometimes swap in memory. Exploratory analyses indicated that incorrectly selected objects on hit trials were closer to the true change than expected by chance, suggesting that ensemble statistics are to some degree informative regarding the approximate location of the change.

Furthermore, the incidence of ensemble-based change detections and guess rate was increased on trials with slow change-detection responses. This could mean that detecting changes based on ensemble representations and guessing takes longer than detecting changes based on individual-item representations. Slow guesses are to be expected if observers resort to guessing only after evaluating whether a change was perceived. The same might apply to ensemble representations if these serve as a backup mechanism to individual-item representations, or it might simply take longer to access or compare information stored in ensemble representations. Alternatively, this finding might indicate that representations of the change degrade with time, thus evoking the false impression of a stronger involvement of (spatially imprecise) ensemble representations. Such degradation would have to occur in a non-spatial manner, as we found no evidence for a smearing-out of incorrect localisation responses across time. The results of the subsequent experiments (see below) also argue rather against temporal degradation.

Experiment 2

Several features of Experiment 1 might have influenced the way participants perform the task. For one, people might get confused by the instruction to select the changed object even if no change was registered. More critically, it is possible that participants knew the correct location whenever they truly detected a change, but forgot it during execution of the change-detection task. For example, recalling the arbitrary key-to-response assignment (which button to press to indicate "change") or processing the feedback (the fixation cross turning green in case of a hit) might have diverted attention away from the changed object and interfered with the maintenance of the change location, or the representation of the change location might degrade passively, as considered above. To test whether any such influence led to an underestimation of localisation performance and to replicate our main finding, we repeated the task, but this time omitted the "change" response and had participants select the changed object immediately.

Methods

A new sample of participants recruited at our university participated in Experiment 2 (n = 20, 1 left-handed, median age: 25 years, range: 20-40 years, 11 female). All participants had normal or corrected-to-normal vision. They gave prior informed consent (in writing) and were paid for their participation. Procedure and design were identical to Experiment 1, except that, instead of first deciding on the presence of a change, participants directly selected the object they believed had changed. They were instructed to click on the fixation cross whenever they had not perceived any change. Furthermore, instead of the CRT screen (1,024 × 768 pixels, 120 Hz) used in Experiments 1 and 3a, we used TFT screens (1,920 × 1,080 pixels, 60 Hz) in Experiments 2 and 3b and adapted the stimulus size (number of pixels) to maintain the perceived stimulus size in terms of degrees of visual angle.

Results

Interpreting the decision to select any object as a "change" response and selecting the fixation cross as a "no-change" response allows calculating the hit rate ($\hat{h}$), false-alarm rate ($\hat{f}$), and true detection rate ($\hat{d}$, Eq. 1) in Experiment 2 even without a dedicated change decision. This assumption appears valid, as none

of the parameters derived accordingly did differ significantly from the respective parameters in Experiment 1 ($\hat{h}$ = .58, $t(36) = 0.64$, $p = .529$, $d = 0.21$, $BF_{01} = 2.70$; $\hat{f}$ = .10, $t(36) = 0.35$, $p = .729$, $d = 0.11$, $BF_{01} = 3.02$; and $\hat{d}$ = .53, $t(36) = 0.56$, $p = .576$, $d = 0.18$, $BF_{01} = 2.80$). Also the capacity estimate according to Pashler's (1988) formula was in the usual range: $\hat{k}_p = 3.19$.

Most importantly, we replicated our main result: the correct object was indicated on only 74% of trials on which an object was selected ($\hat{l}|hit = .74$; i.e., trials on which participants would probably have pressed "change" in the typical task), which is not different from $\hat{l}|hit = .71$ in Experiment 1, $t(36) = 0.80$, $p = .428$, $d = 0.26$, $BF_{01} = 2.46$, but significantly smaller than predicted by a pure-item model ($\frac{\hat{d}}{\hat{h}} = .91$), $t(19) = 9.29$, $p < .001$, $d_z = 2.08$, , $BF_{10} > 76 \times 10^4$. Again, incorrect localisations were somewhat closer to the changed object than expected by chance (mean ordinal distance: 2.67), $t(19) = 5.04$, $p < .001$, $d = 1.13$, $BF_{10} > 369$.

Discussion

In a version of the task in which participants were directly prompted to choose the changed object, we replicated the critical findings from Experiment 1: There were a considerable number of trials on which participants detected a change but were not able to localise it, probably because this detection was based on some kind of ensemble representation or 'sensing'. Thus, in Experiment 1, this critical finding was unlikely due to interference or withdrawal of attention related to performing the (explicit) change-detection response first, or due to temporal degradation of information as to the change location. Also, potential confusion arising from the demand of selecting an object after a "no-change" response cannot explain our results. In fact, given that participants had to guess the location also on trials on which they correctly guessed that a change was present (an estimated 10% of trials) and on trials on which they detected the change based on an ensemble representation, the requirement to select objects on miss trials in Experiment 1was unlikely to be confusing. Furthermore, incorrect localisations were again closer to the true location of the change than expected by chance.

Experiments 3a and 3b

Imperfect localisation performance on hit trials of a standard colour-change-detection task as observed in Experiments 1 and 2 has been reported before by Pailian and Halberda (2015, Exp. 3). However, having a different focus (estimating and improving the reliability of the change-detection task), they did not take differential guessing rates into account and did not examine the implications of this finding for estimates of *k*. Interestingly, they used the localisation component as a means to make participants focus more on individual items. Accordingly, the mere presence of a localisation component might increase the incidence of item-based change detections, so that we might have underestimated the degree to which standard change-detection tasks (and thus estimates of VWM capacity) are affected by ensemble representations.

More generally, it seems plausible that item and ensemble representations might draw from the same pool of resource (Alvarez, 2011, p. 128; Brady et al., 2011, p. 9; Brady & Tenenbaum, 2013, p. 104). If this is the case, participants in Experiments 1 and 2 might have been forced to trade-off between both types of representation: Item representations are needed to perform the localisation task, whereas ensemble representations might be more efficient for performing the change-detection task. To examine whether trade-offs in either direction have distorted the results of Experiments 1 and 2, we ran two control experiments: a pure (i.e., standard) change-detection task (Experiment 3a) and a pure localisation task (Experiment 3b).

Methods

New samples of participants recruited at our university participated in Experiment 3a (n = 16, 1 left-handed, median age: 21.5 years, range: 19-38 years, 12 female) and Experiment 3b (n = 20, 1 left-handed, median age: 24.5 years, range: 18-40 years, 16 female). One data set from Experiment 3a was lost due to technical problems. All participants had normal-or-corrected to normal vision. They gave prior informed consent (in writing) and received course credit or were paid for their participation. Procedure and design were identical to Experiment 1, except that the localisation response was omitted in Experiment 3a and

that a change occurred on all trials and no change-detection response had to be given in Experiment 3b (see Fig. 2 for a comparison of all tasks).

Results

Hit rates were (numerically) somewhat higher and false-alarm rates somewhat lower in Experiment 3a compared to Experiment 1 ($\hat{h}$ = .64, $t(31) = 0.65$, $p = .518$, $d = 0.23$, $BF_{01} = 2.54$; and $\hat{f}$ = .14, $t(31) = 1.91$, $p = .065$, $d = 0.67$, $BF_{10} = 1.31$, respectively) and to Experiment 2 ($t(33) = 1.38$, $p = .177$, $d = 0.47$, $BF_{01} = 1.47$, and $t(33) = 2.42$, $p = .021$, $d = 0.83$, $BF_{10} = 2.86$, respectively), indicating that participants were a little more prone to guessing. Importantly, however, the guessing-corrected detection rate (Eq. 1) was virtually identical to Experiment 1, $\hat{d}$ = .58, $t(31) = 0.30$, $p = .768$, $d = 0.10$, $BF_{01} = 2.90$, and Experiment 2, $t(33) = 0.88$, $p = .386$, $d = 0.30$, $BF_{01} = 2.26$. Also, the capacity estimate according to Pashler's (1988) formula was in the usual range again, $\hat{k}_p$ = 3.49. This indicates that adding a localisation task in the main experiments did not draw resource away from the standard change-detection task; that is, participants did not trade-off ensemble representations for individual-item representations.

There is no estimate of localisation accuracy on hit trials ($\hat{l}|\hat{h}$) in Experiment 3b, because no change-detection response was given. We can still compare performance to that of Experiment 1, because in the latter, participants were asked to select the changed object independently of whether or not they detected a change. Overall localisation accuracy was virtually equivalent ($\hat{l} = .56$ vs. $\hat{l} = .54$), $t(36) = 0.73$, $p = .473$, $d = 0.24$, $BF_{01} = 2.58$. This indicates the change-detection task in the main experiments did not draw resource away from the localisation task; that is, participants did not trade-off individual-item representations for ensemble representations. Again, incorrect localisations were closer to the changed object than expected by chance (mean ordinal distance: 2.81), $t(19) = 6.38$, $p < .001$, $d = 1.43$, $BF_{10} > 5007$.

Discussion

Experiments 3a and 3b provide evidence that performance in the critical Experiments 1 and 2 was not contaminated by trade-offs between item and ensemble representations. That is, participants asked to

detect and localise a change do not appear to trade-off resource between the two tasks (or develop some idiosyncratic strategies that harm performance on any of the sub-tasks). Instead, it appears that, on a certain percentage of trials, location information comes at no cost – as would be expected if these detections are based on item representations that convey location information by default. Additionally, the comparable performance in Experiment 1 (50% changes) and Experiment 3b (100% changes) indicates that uncertainty regarding the occurrence of changes did not influence localisation performance.

General Discussion

We developed and evaluated a new method of estimating the incidence of ensemble-based comparisons in the change-detection task by asking participants to additionally localise the change. The results indicate that even in the most basic version of the task – commonly employed to measure the number of items a person can maintain in VWM (while minimizing the influence of grouping-based ensemble representations by avoiding item repetitions) – change-detection performance is actually influenced by ensemble representations to a considerable degree.

Several control experiments showed that this observation is not attributable to task demands additionally introduced by the new method. In fact, as can be seen by comparing values within rows in Table 2, performance was virtually the same independently of whether the task required pure change detection (Exp. 3a), pure localisation (Exp. 3b), or a mixture of the two (Exp. 1 and 2). Thus, we can conclude that, in the present task, the additional requirement to localise the change likely did not influence change-detection performance, or vice versa.

Beyond the main finding that ensemble representations may well influence performance even in the simplest version of the change-detection task, which is commonly employed to measure VWM capacity, our findings have some additional implications: (a) the ensemble representations employed here might convey some spatial information; (b) there appeared to be no trade-off between individual-item and ensemble representations in the present study; (c) new models, incorporating spatial ensemble representations, for change-detection and change-localisation are needed; and (d) VWM capacity in terms of a fixed number of stored items (*k* index) might typically be overestimated. Before discussing these

implications in turn, we will address a number of objections that might be levelled against our central assumption that item-based change detections allow localising the changed object.

Does Item-Based Change Detection Necessarily Provide Spatial Information?

There are several reasons to assume that any type of item-based change detection necessarily involves information on the location of the change. It is hard to imagine how detecting a change in a specific object might be achieved without also knowing which object has changed. Furthermore, formal mathematical models widely employed to estimate VWM capacity assume that items are compared pairwise based on their spatial location (i.e., the colour patch in, say, the upper left corner of the memory display is compared to the patch in the upper left corner of the test display, but not to the patch in the lower right corner of the test display; Pashler, 1988; Rouder et al., 2011) – that is, spatial item position is considered an inherent part of the comparison process. Others have provided evidence that changes attract spatial attention (Hyun, Woodman, Vogel, Hollingworth, & Luck, 2009). Accordingly, when attention is attracted automatically by a change, participants would merely have to maintain attention, or the eye, at the change position to make an explicit localisation response later on.

Non-spatial item-based comparisons. Even if (some) comparisons were non-spatial, people could still easily locate the change by selecting the object in the test array that has the colour for which the comparison failed. In Oberauer and Lin's (2016) model of VWM, for example, observers maintain not only feature bindings (e.g., colour-location bindings) but also pure, context-free features (e.g., something akin to a [non-verbal] list of colours). Accordingly, change detection may sometimes be based on a mismatch of the binding ("there was another colour at that position"; which would require maximally n comparisons, with n = number of objects) and sometimes on a mismatch of the feature ("this colour was not present before"; which would require maximally n^2 comparisons). In case of binding mismatches, people must have information on the location, because the location defines the binding (and the pairing of objects for comparison). But even with a pure feature mismatch they can easily localise the change by identifying the colour that was not present before (for which the pairwise matching failed) in the test display. For example, if they notice that there is a blue square in the test array and blue was not present in

the memory array, the blue square must be the changed object. Thus, the information on the location of the change could be retrieved from the test display and must therefore not necessarily be maintained to accurately perform change localisation.

No conscious access. One might claim that even though the comparison process is location-based, participants do not have conscious access to this information. First, conscious access is not necessary to influence behaviour (see, e.g., Newell & Shanks, 2014, and the ensuing commentaries). Second, the very notion of VWM often goes along with (access) consciousness (Block, 2011; Dehaene & Naccache, 2001; Velichkovsky, 2017; but see Soto & Silvanto, 2016). Finally, in the majority of trials our participants were able to localise the change, thus one would have to assume that the location information is sometimes consciously accessible and sometimes not, due to some yet undetermined factors. Conscious vs. unconscious change-detection is equally (non-)parsimonious as item-representation vs. ensemble-representation change detection and seems less supported by current theories and available data.

Rapid forgetting of the change location. A more plausible reason to assume a mixture of detection with and without localisation within a pure-item model might be that information on the changed object is initially consciously accessible when the test display appears, but then rapidly forgotten, so that it can be used only on some of the trials. This would be in line with the estimated higher incidence of spatially imprecise ensemble representations on trials with slow change-detection responses in Experiment 1, which might simply reflect a decrease in spatial precision of the change representation as time passes. However, if there was such a gradual degradation, more of this spatial information should have been forgotten in Experiment 1 than in Experiments 2 or 3b, because participants were allowed directly to put the information to use in Experiment 2 and 3b, whereas they had to defer the localisation response until after the change-detection response and the corresponding feedback in Experiment 1. Nevertheless, it remains a theoretical possibility that degradation of the change representation maxed out near-instantaneously, so that initial knowledge of the location of the change is underestimated even with immediate localisation responses (Experiments 2 and 3b).

Spatial Ensemble Representations?

The ensemble representations influencing change detection posited by Brady and colleagues (Brady & Alvarez, 2015b; Brady & Tenenbaum, 2013) contain spatial information. Brady and Tenenbaum's (2013) model, for example, assumes that VWM makes use of redundancies in the memory display by forming spatial ensemble representations of similar, neighbouring objects, akin to data-compression algorithms (see also Fig. 1A-C). If a change in such an ensemble is detected and observers are asked to localise it, they might be able to constrain localisation to the sub-group of objects that are represented by the ensemble representation, rather than choosing an item from any other sub-group. When items within a sub-group containing the change are preferably selected, localisation responses are (on average) closer to the true change than expected by chance. Therefore, the observed presence of a spatial bias in erroneous selections may be indicative of the formation of such sub-groups, whereas its absence would have been indicative of ensemble representations spanning all objects in the display. Given that we explicitly avoided feature repetitions and used clearly delineable colours, groupings based on featural similarity were less likely in the present study (although future studies might shed light on whether certain non-identical, but similar features are more likely to be grouped; see Son, Oh, Kang, & Chong, 2018). It appears reasonable to speculate that our observers formed ensemble representations of sub-groups based on the spatial arrangement of the display alone (e.g., 'average' colour of several nearby objects), thus producing the observed spatial bias in erroneous localisations. However, deciding whether ensemble representations are responsible for the observed spatial bias requires further, dedicated research and direct comparison with alternative models that might produce such a spatial bias based on individual-item representations, perhaps by assuming an increased rate of swapping between nearby objects (see Appendix B).

No Trade-Off Between Individual-Item and Ensemble Representations?

Given that change detection can be achieved based on individual-item representations and, respectively, on ensemble representations, participants might strategically choose an ideal trade-off between both types of representation. A trade-off would be necessary if individual-item representations and ensemble representations draw from the same limited-capacity resource (e.g., the overall amount of a

flexibly allocable resource or the number of available slots); that is, when ensemble representations "take up space in memory that would otherwise be used to represent more information about individual items" (Brady & Tenenbaum, 2013, p. 104; see also Alvarez, 2011, p. 128; Brady et al., 2011, p. 9).

To gage how our data might inform this question, consider that ensemble representations would typically be less useful for change localisation than for change detection, because they provide information about whether a change occurred, but only very rough information (if any) about where the change occurred. Consequently, one might assume that all resource was allocated to representing individual items in Experiment 3b, in which participants had to select the changed item and a change occurred on every trial. As no change decision was needed, the information potentially gained from ensemble representations would have limited value for solving this specific task. Given this, the finding of equal localisation performance in Experiment 3b and Experiment 1(in which an explicit change decision was required) might be taken to suggest that these representations do not draw on the same resource; in other words, ensemble representations may provide extra capacity.

On the other hand, our data are also reconcilable with the assumption that resource is shared between ensemble representations and individual-item representations, though with certain constraints. In particular: the pure change-detection task (Exp. 3a) might already motivate a maximization of individual-item representations, but at least one ensemble representation may have to be hold mandatorily. Under different task conditions, participants might decide to encode the displays at higher levels of abstraction, thus moving from representations of single features (individual items) to representations containing multiple features (i.e., they could decide to strategically use more ensemble-like representations; e.g., Greene & Oliva, 2009; Nie et al., 2017).

Overestimations of *k*

We found that even in a simple standard task widely used to measure *k* (see the list of recent publications in Appendix A) and even without any feature repetitions (see Brady & Tenenbaum, 2013), ensemble representations did contribute to change-detection performance. As ensemble representations are not taken into account in the pure-item models underlying the estimation of *k*, VWM capacity might be

routinely overestimated. To gain a quantitative impression of this overestimation of *k*, we developed a respective mathematical model and applied it to our data (Appendix A). See Table 3 for the results. For comparison, the table also reports k_p according to the traditional pure-item model (Pashler, 1988; Rouder et al., 2011).

Of note, compared to pure-item models, estimates of *k* decrease from $\widehat{k_p} = 3.34$ (averaged across Experiments 1, 2, and 3a) to $\widehat{k^*} = 2.30$ (averaged across Experiments 1 and 3b; with k^* being the upper bound on true *k*). Thus, *k* values from an item-plus-ensemble model are more than 31% (1.04 items) lower than indicated by standard formulae of *k* (Pashler, 1988; Rouder et al., 2011). That this is a substantial difference can be seen from a comparison with an effect on *k* that is considered "very large" according to an influential review paper (Luck & Vogel, 2013, p. 397): in one study, mean *k* values differed between schizophrenics (2.34) and healthy controls (2.93) by 0.59 items (d = 1.11, Johnson et al., 2013).

Future Directions

Given this overestimation of *k* based on pure-item models, it might seem surprising that precisely this *k* correlates highly with neuronal indices of the number of items maintained in VWM (Todd & Marois, 2004; Vogel & Machizawa, 2004; Xu & Chun, 2006) and with general cognitive functions such as intelligence (Fukuda et al., 2010; Unsworth, Fukuda, Awh, & Vogel, 2014). However, these correlations are not perfect ($r < 1$); in fact, a meta-analysis of the correlation between *k* and the most often employed neurophysiological measure of VWM capacity (contralateral delay activity, CDA; Vogel & Machizawa, 2004) revealed a correlation of maximally $r = .67$ (upper bound of the 95% confidence interval; Luria et al., 2016). That is, at least $1 - r^2 = 55\%$ of variance in *k* is not explained by the CDA, quite possibly in part because change-detection performance, but not the CDA, is influenced by ensemble representations. Another interesting question for future studies would be to determine the degree to which item- versus ensemble-based change-detection performance predicts intelligence (Fukuda et al., 2010; Unsworth, Fukuda, Awh, & Vogel, 2014).

In the present, initial study, we examined whether localisation performance could be used to demonstrate the involvement of ensemble-based change detections in the simplest, standard version of the

change-detection task. There might, of course, be several factors that determine the relative influence of the two, item- and ensemble-based representations. For example, longer encoding times might yield a higher incidence of item-based change detections, because observers have more time to 'home in' on individual items; however, ensemble representations are known to increase in quality with encoding time (e.g., Whitney & Yamanashi Leib, 2018), so that observers might alternatively chose to more heavily rely on these. Furthermore, one or the other representation might degrade more easily during the retention interval, so that the length of this interval might modulate the relative incidence of item- and ensemble-based change detections (e.g., Pertzov, Bays, Joseph, & Husain, 2013; Pertzov, Manohar, & Husain, 2017; Souza, Rerko, & Oberauer, 2016).

Interestingly, in Experiment 1, incorrectly selected objects were numerically slightly closer to the change than expected by chance, even on miss trials (see Fig. 3). Although the respective test was non-significant and resulted in an indecisive Bayes factor of $BF_{10} = 1.21$, it might be informative to consider what this would mean if it was a real effect. Notably, on the hypothesis that the spatial bias in erroneous localisation responses is driven by ensemble representations, this might indicate that observers sometimes detected a change based on the ensemble representation, but still did not press 'change' – a potential reason being that they were not sufficiently confident that anything had changed, but when forced to choose one object, they still utilised the available spatial information to a certain degree. Future studies revisiting this specific issue might benefit from including confidence ratings.

Another interesting question is whether and how irrelevant objects in the memory display that do not need to be remembered (*distractors*; e.g., Feldmann-Wüstefeld, & Vogel, 2018; Liesefeld, Liesefeld, & Zimmer, 2014; Vogel, McCollough, & Machizawa, 2005) might be incorporated into ensemble representations, so that observers would sometimes erroneously select distractor locations after an ensemble-based change detection. The exact reasons for the spatial bias in erroneous localisations requires further research in any case and might yield interesting insights into the nature of ensemble representations employed in change detection.

References

Agostinelli, G., Sherman, S. J., Fazio, R. H., & Hearst, E. S. (1986). Detecting and identifying change: Additions versus deletions. Journal of Experimental Psychology: Human Perception and Performance, 12, 445-454. doi:10.1037/0096-1523.12.4.445

Allon, A. S., & Luria, R. (2017). Compensation mechanisms that improve distractor filtering are short-lived. Cognition, 164, 74-86. doi:10.1016/j.cognition.2017.03.020

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. Journal of Experimental Child Psychology, 106, 20-29. doi:10.1016/j.jecp.2009.11.003

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15,* 122-131. doi:10.1016/j.tics.2011.01.003

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*, 106-111. doi:10.1111/j.0963-7214.2004.01502006.x

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *PNAS, 106,* 7345-7350. doi:10.1073/pnas.0808981106

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*, 622-628. doi:10.1111/j.1467-9280.2007.01949.x

Ball, F., & Busch, N. A. (2015). Change detection on a hunch: Pre-attentive vision allows 'sensing' of unique feature changes. *Attention, Perception, & Psychophysics*, *77*, 2570-2588. doi:10.3758/s13414-015-0963-9

Bargh, J. A. (2016). Awareness of the prime versus awareness of its influence: Implications for the real-world scope of unconscious higher mental processes. *Current Opinion in Psychology*, *12*49-52. doi:10.1016/j.copsyc.2016.05.006

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *The Journal of Neuroscience, 34,* 3632-3645. doi:10.1523/JNEUROSCI.3204-13.2014

Bays, P. M. (2015). Spikes not slots: Noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*, 431-438. doi:10.1016/j.tics.2015.06.004

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*, 851-854. doi:10.1126/science.1158023

Bays, P. M., & Schneegans, S. (2019). New perspectives on binding in visual working memory. *British Journal of Psychology* [Manuscript in press]

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*, 1622-1631. doi:10.1016/j.neuropsychologia.2010.12.023

Becker, M. W., Pashler, H., & Anstis, S. M. (2000). The role of iconic memory in change-detection tasks. *Perception*, *29*, 273-286. doi:10.1068/p3035

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, *15*, 567-575. doi:10.1016/j.tics.2011.11.001

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*, 384-392. doi:10.1177/0956797610397956

Brady, T. F., & Alvarez, G. A. (2015a). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision*, *15*(15), 6. doi: 10.1167/15.15.6.

Brady, T. F., & Alvarez, G. A. (2015b). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 921-929. doi:10.1037/xlm0000075

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*, 487-502. doi:10.1037/a0016797

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 4, doi:10.1167/11.5.4

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review, 120*, 85-109. doi:10.1037/a0030779

Busch, N. A., Dürschmid, S., & Herrmann, C. S. (2010). ERP effects of change localization, change identification, and change blindness. *NeuroReport, 21,* 371-375. doi:10.1097/WNR.0b013e3283378379

Chong, S. C., Joo, S. J., Emmanouil, T., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics, 70,* 1327-1334. doi:10.3758/PP.70.7.1327

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43,* 393-404. doi:10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45,* 891-900. doi:10.1016/j.visres.2004.10.004

Conway, A. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences, 7,* 547-552. doi:10.1016/j.tics.2003.10.005

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87-185. doi:10.1017/S0140525X01003922

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51,* 42-100. doi:10.1016/j.cogpsych.2004.12.001

Curby, K. M., Smith, S. D., Moerel, D., & Dyson, A. (2019). The cost of facing fear: Visual working memory is impaired for faces expressing fear. *British Journal of Psychology* [Manuscript in press]

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review, 17,* 673-679. doi:10.3758/17.5.673

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79,* 1-37. doi:10.1016/S0010-0277(00)00123-2

D'Errico, J. (2006). fminsearchbnd.m [Matlab code]. Retrieved from https://github.com/Data2Dynamics/d2d/tree/master/arFramework3/ThirdParty/FMINSEARCHBND

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12), 10, doi:10.1167/11.12.3

Fougnie, D., Cormiea, S. M., & Alvarez, G. A. (2013). Object-based benefits without object-based representations. *Journal of Experimental Psychology: General*, *142*, 621-626. doi:10.1037/a0030300

Feldmann-Wüstefeld, T., & Vogel, E. K. (2018). Neural Evidence for the contribution of active suppression during working memory filtering. *Cerebral Cortex*. Advance online publication. doi:10.1093/cercor/bhx336

Gold, J. M., Robinson, B., Leonard, C. J., Hahn, B., Chen, S., McMahon, R. P., Luck, S. J. (2018). Selective attention, working memory, and executive function as potential independent sources of cognitive dysfunction in schizophrenia, *Schizophrenia Bulletin*. Advance online publication. doi:10.1093/schbul/sbx155

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*, 137-176. doi:10.1016/j.cogpsych.2008.06.001

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, *18*, 855-859. doi:10.3758/s13423-011-0125-6

Heinz, A. J., & Johnson, J. S. (2017). Load-dependent increases in delay-period alpha-band power track the gating of task-irrelevant inputs to working memory. *Frontiers in Human Neuroscience*, *11*, 250. doi:10.3389/fnhum.2017.00250

Hughes, H. C., Caplovitz, G. P., Loucks, R. A., Fendrich, R., & Hamed, S. B. (2012). Attentive and pre-attentive processes in change detection and identification. *Plos ONE*, *7*(8), 1-15. doi:10.1371/journal.pone.0042851

Hyun, J., Woodman, G. F., Vogel, E. K., Hollingworth, A., & Luck, S. J. (2009). The comparison of visual working memory representations with perceptual inputs. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 1140-1160. doi:10.1037/a0015019

Jarmasz, J., & Hollands, J. G. (2009). Confidence intervals in repeated-measures designs: The number of observations principle. *Canadian Journal of Experimental Psychology, 63,* 124–138. doi:10.1037/a0014164

Johnson, M. K., McMahon, R. P., Robinson, B. M., Harvey, A. N., Hahn, B., Leonard, C. J., & ... Gold, J. M. (2013). The relationship between working memory capacity and broad measures of cognitive ability in healthy adults and people with schizophrenia. *Neuropsychology, 27,* 220-229. doi:10.1037/a0032060

Kornblith, S., Buschman, T. J., & Miller, E. K. (2016). Stimulus load and oscillatory activity in higher cortex. *Cerebral Cortex, 26,* 3772-3784. doi:10.1093/cercor/bhv182

Li, S., Cai, Y., Liu, J., Li, D., Feng, Z., Chen, C., & Xue, G. (2017). Dissociated roles of the parietal and frontal cortices in the scope and control of attention during visual working memory. *NeuroImage, 149,* 210-219. doi:10.1016/j.neuroimage.2017.01.061

Liesefeld, A. M., Liesefeld, H. R., & Zimmer, H. D. (2014). Intercommunication between prefrontal and posterior brain regions for protecting visual working memory from distractor interference. *Psychological Science, 25,* 325-333. doi:10.1177/0956797613501170

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490. doi:10.3758/BF03210951

Luck, S. J. (2008). Visual short-term memory. In S. J. Luck & A. Hollingworth (Eds.), Visual Memory (pp. 43-85). New York, NY: Oxford University Press.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279-281. doi:10.1038/36846

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*, 391-400. doi:10.1016/j.tics.2013.06.006

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*, 347-356. doi:10.1038/nn.3655

Mathias, S. R., Knowles, E. E., Barrett, J., Beetham, T., Leach, O., Buccheri, S., & ... Glahn, D. C. (2017). Deficits in visual working-memory capacity and general cognition in african americans with psychosis. *Schizophrenia Research*. Advance online publication. doi:10.1016/j.schres.2017.08.015

Morey, R. D. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology*, *55*, 8-24. doi:10.1016/j.jmp.2010.08.008

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*, 1-18. doi:10.1017/S0140525X12003214

Nie, Q., Müller, H. J., & Conci, M. (2017). Hierarchical organization in visual working memory: From global ensemble to individual object structure. *Cognition*, *159*, 85-96. doi:10.1016/j.cognition.2016.11.009

Oberauer, K., & Lin, H. (2016). An interference model of visual working memory. *Psychological Review*, *124*, 21-59. doi:10.1037/rev0000044

Pailian, H., & Halberda, J. (2015). The reliability and internal consistency of one-shot and flicker change detection for measuring individual differences in visual working memory capacity. *Memory & Cognition*, *43*, 397-420. doi:10.3758/s13421-014-0492-0

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, *44*, 369-378. doi:10.3758/BF03210419

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (2013). Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 1224-1231. doi:10.1037/a0030947

Pertzov, Y., Manohar, S., & Husain, M. (2017). Rapid forgetting results from competition over time between items in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 528-536. doi:10.1037/xlm0000328

Rensink, R. A. (2004). Visual sensing without seeing. *Psychological Science*, *15*, 27-32. doi:10.1111/j.0963-7214.2004.01501005.x

Robison, M. K., McGuirk, W. P., & Unsworth, N. (2017). No evidence for enhancements to visual working memory with transcranial direct current stimulation to prefrontal or posterior parietal cortices. *Behavioral Neuroscience*, *131*, 277-288. doi:10.1037/bne0000202

Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin & Review*, *18*, 324-330. doi:10.3758/s13423-011-0055-3

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225-237. doi:10.3758/PBR.16.2.225

Shin, H., & Ma, W. J. (2017). Visual short-term memory for oriented, colored objects. *Journal of Vision*, *17*, doi:10.1167/17.9.12

Simmering, V. R., & Wood, C. M. (2017). The development of real-time stability supports visual working memory performance: Young children's feature binding can be improved through perceptual structure. *Developmental Psychology, 53*, 1474-1493. doi:10.1037/dev0000358

Son, G., Oh, B., Kang, M., & Chong, S. C. (2018, May). *Similarity-based clusters are the representational units of visual working memory.* Poster presented at the Vision Science Society 18th Annual Meeting, St. Pete Beach, FL.

Soto, D., & Silvanto, J. (2016). Is conscious awareness needed for all working memory processes? *Neuroscience of Consciousness, 1*, niw009. doi:10.1093/nc/niw009

Souza, A. S., Rerko, L., & Oberauer, K. (2016). Getting more from visual working memory: Retro-cues enhance retrieval and protect from visual interference. *Journal of Experimental Psychology: Human Perception and Performance, 42,* 890-910. doi:10.1037/xhp0000192

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428,* 751-754. doi:10.1038/nature02466

Turatto, M., & Bridgeman, B. (2005). Change perception using visual transients: Object substitution and deletion. *Experimental Brain Research, 167,* 595-608. doi:10.1007/s00221-005-0056-4

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology, 71,* 1-26. doi:10.1016/j.cogpsych.2014.01.003

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121,* 124-149. doi:10.1037/a0035234

van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, 109,* 8780-8785. doi:10.1073/pnas.1117465109

Velichkovsky, B. B. (2017). Consciousness and working memory: Current trends and research perspectives. *Consciousness and Cognition, 55,* 35-45. doi:10.1016/j.concog.2017.07.005

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature, 428,* 748-751. doi:10.1038/nature02447

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature, 438,* 500-503. doi:10.1038/nature04171

Weaver, M. D., Hickey, C., & van Zoest, W. (2017). The impact of salience and visual working memory on the monitoring and control of saccadic behavior: An eye-tracking and EEG study. *Psychophysiology, 54,* 544-554. doi:10.1111/psyp.12817

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General, 131,* 48-64. doi:10.1037/0096-3445.131.1.48

Wijeakumar, S., Magnotta, V. A., & Spencer, J. P. (2017). Modulating perceptual complexity and load reveals degradation of the visual working memory network in ageing. *NeuroImage, 157*, 464-475. doi:10.1016/j.neuroimage.2017.06.019

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology, 69*, 105-129. doi:10.1146/annurev-psych-010416-044232

Xie, W., & Zhang, W. (2017). Discrete item-based and continuous configural representations in visual short-term memory. *Visual Cognition, 25*, 21-33. doi:10.1080/13506285.2017.1339157

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature, 440*, 91-95. doi:10.1038/nature04262

Appendix A: An item-plus-ensemble model of change detection and change localisation

Preliminary remarks. Appendix A explores the implication of an involvement of ensemble representations in change detection on the most common estimate of VWM capacity, namely *k*. By taking *k* literally as the number of items that can be stored in VWM, one (implicitly) accepts the notion of a limited number of concurrent representations in VWM (*slots*; Cowan, 2001; Luck & Vogel, 2013; Morey, 2011; Pashler, 1988; Rouder et al., 2011). Measuring VWM capacity as some fixed number of representations has, therefore, only pragmatic (if any) value for those who believe in an infinitely divisible VWM resource (Bays, 2014, 2015; Bays & Husain, 2008; Ma et al., 2014; van den Berg et al., 2012, 2014). However, even if the discrete-slot assumption turns out to be false, there are still several reasons why a reconsideration of *k* is of relevance.

For one, there is a reasonable interpretation for *k* even in continuous-resource theories: To explain change-detection performance, such models must assume that a change-response is triggered when evidence for a change surpasses a certain threshold – where this degree of evidence would likely depend on the precision of the representation of the changed item. Accordingly, *k* can be re-interpreted as the average number of items that can be maintained with sufficient precision to detect a change in the specific task used. As more items can attain this precision when more resource is available, *k* would be a direct correlate of the total amount of available resource, and this correlate could potentially be biased by the involvement of ensemble representations.

Also note that the interpretation of the detection rate *d* is not influenced by the nature of the capacity limitation, because it is agnostic as to what caused the change decision. Thus, if there was an additional influence on *d* beyond individual-item representations, this would be of relevance for discrete-slots and continuous-resource models alike.

There are, of course, also purely pragmatic reasons for an interest in *k*: it has a huge appeal due to its utility as an explanatory construct for understanding individual differences in cognitive functioning and for relating change-detection performance to neuronal markers of VWM maintenance (see General Discussion; e.g., Fukuda et al., 2010; Todd & Marois, 2004; Vogel & Machizawa, 2004; Vogel et al.,

2005; Xu & Chun, 2006). Furthermore *k* is still heavily used in various psychological disciplines (for some recent examples, see, Allon & Luria, 2017; Curby, Smith, Moerel, & Dyson, 2019; Heinz & Johnson, 2017; Li et al., 2017; Mathias et al., 2017; Nie, Müller, & Conci, 2017; Robison, McGuirk, & Unsworth, 2017; Simmering & Wood, 2017; Weaver, Hickey, & van Zoest, 2017; Wijeakumar, Magnotta & Spencer, 2017; Xie & Zhang 2017), demonstrating that (a) slot-models are not abandoned by the research community and (b) there is a pragmatic need to re-evaluate the appropriateness of the assumptions underlying estimates of *k* as new knowledge on the nature of change detection is being generated.

Model. Similar to prior, pure-item models (Pashler, 1988; Rouder et al., 2011), we assume that correct "change" responses (hits) are either due to a correct detection of a change (informed by either type of representation) or due to a lucky guess when there was no true detection. Thus, hit rate (h) is determined by true detection rate (d) and guessing rate (g). The latter (the proneness to guess when in fact no change was detected) can be estimated from the incidence of trials on which we know that participants did not truly detect a change, because there was none, but answered 'change' nonetheless (i.e., the false-alarm rate, f). In particular,

$$h = d + (1 - d) \cdot g \quad (2)$$

and

$$f = g \quad (3)$$

Substituting f for g and solving for d yields Pashler's formula (Pashler, 1988; Rouder et al., 2011):

$$h = d + (1 - d) \cdot f [= d + f - df]$$

$$\Leftrightarrow h - f = d - df [= d \cdot (1 - f)]$$

$$\Leftrightarrow d = \frac{h - f}{1 - f} \quad (4)$$

In contrast to Pashler (1988) and Rouder et al. (2011), we assume that true detections (d) are either item-based ($d_i = \frac{k}{S}$) or ensemble-based (d_e), with k^* = number of *item* representations VWM can hold and S = set size (number of items in the memory display), thus

$$d = d_i + d_e = \frac{k^*}{S} + d_e \quad (5)$$

Substituting in (4) and solving for k^* yields

$$k^* = \left(\frac{h-f}{1-f} - d_e\right) \cdot S \quad (6)$$

As there is an unknown on the right side of (6), namely d_e, we can no longer use this Equation to determine *k*. In general, averaged change-detection data is insufficient to determine *k* if an influence of ensemble representations on change-detection performance is granted and, as d_e is positive, the commonly employed formula for *k* (Pashler, 1988; Rouder et al., 2011) will overestimate the number of individual item representations in VWM (i.e., it will overestimate VWM capacity).

To determine *k* given the influence of ensemble representations on change detection, additional data are needed, as, for example, that from the localisation task. Two outcomes of the present studies pose helpful restrictions on the theoretically possible model space for localisation performance. First, even on miss trials (when participants had not detected any change), localisation performance was clearly above chance (i.e. $> \frac{1}{S}$) in Experiment 1. This means that guessing was informed (rather than being completely random). The model will therefore assume that when no change is perceived, guessing the change location will be restricted to those objects that were not individually represented in memory. In other words, participants would not select an object they had in memory and therefore knew that it did not change. Second, the presence of a spatial bias in erroneous localisations might indicate that the employed ensemble representations contain imprecise information on the location of the change and, therefore, further restrict the set of potential change locations, thus further informing guessing.

Thus, we assume that correct localisations (l) are driven by item representations and informed guessing as just described. In particular, when the specific changed item has been represented (which is

the case in $\frac{k^*}{S}$ of trials), participants will select the correct location. If we could be sure that there is no spatial information in the employed ensemble representations, the rest would be easy: whenever the changed item was not in VWM (i.e., on $1-\frac{k^*}{S}$ of trials), participants would select at chance an item from among those that were not individually represented (informed guessing) and thus be correct on $\frac{1}{S-k^*}$ of cases. With these assumptions, we could derive a formula for k^* based on localisation performance (see also van den Berg et al., 2012, Supporting Information, p. 2)

$$l = \frac{k^*}{S} + \left(1-\frac{k^*}{S}\right)\cdot\frac{1}{S-k^*} = \frac{k^*}{S} + \frac{S-k^*}{S}\cdot\frac{1}{S-k^*} = \frac{k^*}{S} + \frac{1}{S} = \frac{k^*+1}{S}$$

$$\Leftrightarrow\ k^* = lS - 1 \quad (7)$$

Rearranging (2) yields

$$d_e = d - \frac{k^*}{S}$$

Entering (3) and (6) into this equation yields a formula for estimating the incidence of ensemble-based change detections:

$$d_e = \frac{h-f}{1-f} - \frac{lS-1}{S} = \frac{h-f}{1-f} - l + S^{-1} \quad (8)$$

However, the additional spatial information potentially conveyed by ensemble representations might further improve guessing to some unknown degree *x* and therefore,

$$l = \frac{k}{S} + \left(1-\frac{k}{S}\right)\cdot\left(\frac{1}{S-k}+x\right) = \frac{k}{S} + \left(1-\frac{k}{S}\right)\cdot\frac{1}{S-k} + \left(1-\frac{k}{S}\right)\cdot x$$

substituting $z = \left(1-\frac{k}{S}\right)\cdot x$ yields:

$$l = \frac{k}{S} + \left(1-\frac{k}{S}\right)\cdot\frac{1}{S-k} + z = \frac{k+1}{S} + z \Leftrightarrow\ k = (l-z)\cdot S - 1 \quad (9)$$

Given that $\left(1-\frac{k}{S}\right)$ and *x* are necessarily positive (one represents a [converse] probability and the other is a performance *benefit*), *z* (their product) is positive as well. Therefore, *k* according to (7) is maximal for *x* = 0 (no localisation benefit due to ensemble representations). In this case, (9) reduces to (7), which consequently is the upper bound on true *k*. (8) then gives the lower bound on the incidence of

ensemble-based change detections (because true *k* is smaller than k^* so that more of the detections are based on ensemble representations than estimated by (8)). Another, less mathematical, way to see that (7) provides the upper bound on *k* is to consider that the additional spatial information potentially conveyed by spatial ensemble representations must always improve localisation performance and, therefore, inflate k^*, so that true *k* is overestimated, because (7) does not take any additional spatial information into account.

Appendix B: Model fitting

The pure-item model, which also constitutes the basis for the other two, more complex, models is (with one addition to predict *l*|*m*; see below) fully described by Equations 2, 3 and 7 in Appendix A. In short, it assumes that a change is detected and localised when the changed item was represented individually in working memory. People also guess on the change-detection part (Eq. 2 and 3) and localisation performance is further enhanced by informed guessing, as detailed above (Eq. 7).

The item-plus-ensemble model additionally assumes that changes can be detected based on a mismatch with the ensemble representation. If s_e is the average number of items included into an ensemble representation, the ensemble-based detection rate can be specified as

$$d_e = \frac{s_e}{S}$$

On the hypothesis that observers make best use of their limited VWM store, the model assumes that only items that are not already contained in the ensemble representation are selected for being represented individually, so that a change is detected either based on an item or on an ensemble representation. With this assumption, (5) remains true and we can use it to solve (2) to predict hit rate.

Neither item nor ensemble representations play a role on no-change trials, so that false-alarm rate is still equal to guessing rate (Eq. 3). Assuming that ensemble representations are involved during localisation would cause various problems (e.g., informed guessing would be perfect with $k = 2$ and $s_e = 3$) and require additional parameters (e.g., reflecting the rate with which ensemble information is used for localisation). For the present purposes, we therefore simply assume that ensemble representations are not involved or negligible during localisation, potentially because their spatial-information content is too unreliable. In this case, localisation rate is given by (7) as well.

Another reason for the difference between hit and localisation rate could theoretically be that items maintained in VWM swap colours with a certain rate (*sw*; Bays, 2016; Bays et al., 2009; Schneegans & Bays, 2019). A swap would produce two (additional) mismatches between memory representation and test display and can thus trigger a “change” response on both change and no-change trials; people would resort to guessing only if they do not perceive any (illusory) change, thus

$$h = d + (1 - d) \cdot sw + (1 - d - sw) \cdot g$$

and

$$f = sw + (1 - sw) \cdot g$$

Crucially, swaps with the changed (target) item would decrease localisation rate without changing hit rate, because, when perceiving two changes, people would have to guess which of the two is the actual change. Swaps between to non-target items would only increase the confusion, because people would perceive three changes. The proportion of such non-target swaps (sw_n) on change trials is given by

$$sw_n = sw \cdot \left(1 - \frac{1}{k}\right) \cdot \left(1 - \frac{1}{k - 1}\right) = sw \cdot \left(1 - \frac{2}{k}\right)$$

The proportion of swaps including the target (sw_t) consequently is

$$sw_t = sw \cdot (1 - sw_n) = sw \cdot \frac{2}{k}$$

Subtracting such erroneous localisations from the detection rate and adding an informed guessing component as in the other models, the predicted localisation rate is

$$l = d - sw_t \cdot \frac{1}{2} - sw_n \cdot \frac{2}{3} + \frac{1 - d - sw}{S - k}$$

Note that the swapping rate is equal on trials on which the target item was and was not in memory, and people necessarily chose the wrong item when a swap occurred on the latter trials, $(1 - d) \cdot sw \cdot 0 = 0$, so that the incidence of guessing is indeed reduced by the overall swapping rate (numerator of last term in the equation).

Finally, localisation rate on miss trials is the same for all three models, because ensemble representations are assumed to be ignored during informed guessing and no swap occurred on "no-change"-response trials,

$$l|m = \frac{1}{S - k}$$

These three models were fit to the empirical data of Experiment 1 (the only Experiment providing a sufficient number of independent data points) using fminsearchbnd.m (Release 4, 7/23/06; D'Errico, 2006) in Matlab (The Mathworks, Natick, MA, USA), with lower bounds = 0 on all parameters and upper

bounds = 1 on all parameters but s_e (upper bound: S) and k (upper bound: inf). Best fitting parameters and model predictions are listed in Table B1 and Table 1, respectively.

Table 1 is discussed in the main text, but Table B1 provides some additional insights that appear worth discussing here. First, k in the pure-item model is lower than k_p, because the fitting algorithm settled on a compromise between the misfit in localisation rate and hit rate. Second, k in the item-plus-ensemble model is exactly k^* (the lower bound on k according to a broader family of item-plus-ensemble models), because in this specific model we assumed that ensemble representations do not contribute to localisation rate, so that only memories of individual items informed localisation (i.e., x and, consequently, z in Eq. 9 is zero). Third, $s_e \approx 1$ might appear surprising, but note that this is the *average* size of ensemble representations. Thus, if, on a given trial, an ensemble representation is employed, it likely contains more than one item. It might indeed always contain all items in the display if ensemble representations are employed on only $1/6^{th}$ of trials (compare this to the respective estimate of $d_e = .17$ in Table 3). Smaller ensemble representations would have to occur on respectively more trials to produce the same average size. Finally, that $g = 0$ and $sw = f$ in the swapping model means that all false alarms are explained by swaps and the assumption of guessing becomes superfluous (cf. van den Berg et al., 2012. This is an interesting feature of the swapping model that seems to warrant further exploration in future studies, using versions of the model that potentially provide a better fit with the empirical data.

Table 1

Empirically Observed and Model-Predicted Performance (Deviations in Brackets)

Parameter	Empirical	Pure-Item	Item+Ensemble	Swapping
$\hat{h}$	.61	.52 (-0.09)	.61 (0.00)	.54 (-0.07)
$\hat{f}$	.10	.10 (0.00)	.10 (0.00)	.10 (0.00)
$\hat{l}$	.56	.63 (0.06)	.56 (0.00)	.56 (0.00)
$\hat{l}\|miss$	.31	.31 (0.00)	.28 (-0.03)	.33 (0.02)

Note. $\hat{h}$ = hit rate; $\hat{f}$ = false alarm rate; $\hat{l}$ = correct-localisation rate. For model details, see Appendix B.

1

Table 2

Summary of Results Across All Experiments

Parameter	Experiment 1	Experiment 2	Experiment 3a	Experiment 3b
$\hat{h}$	.61	.58	.64	-
$\hat{f}$	.10	.10	.14	-
$\hat{d}$	.56	.53	.58	-
$\hat{l}$	.56	-	-	.54
$\hat{l}\|hit$	.71	.74	-	-
$\hat{l}\|miss$	.31	-	-	-

Note. $\hat{h}$ = hit rate; $\hat{f}$ = false alarm rate; $\hat{d}$ = true detection rate (Eq. 1); $\hat{l}$ = correct-localisation rate. Given the different types of responses, different pieces of information are extractable from the four experiments. However, comparisons within lines indicate that performance was remarkably stable across the different tasks and subject samples.

Table 3

Parameter Estimates From the Proposed Item-Plus-Ensemble Model ($\hat{d}_e, \hat{d}_i, and\ \hat{k}^*$) and From the Standard Pure-Item Model ($\hat{k}_p$; Rouder et al., 2011; Pashler, 1988.)

Parameter	Experiment 1	Experiment 2	Experiment 3a	Experiment 3b
$\hat{d}_e$	.17	-	-	-
$\hat{d}_i$	.40	-	-	-
$\hat{k}^*$	2.39	-	-	2.23
$\hat{k}_p$	3.38	3.19	3.49	-

Note. **item-plus-ensemble model:** $\hat{d}_e$ = lower bound on detection rate based on ensemble-representations (Eq. 5); $\hat{d}_i$= upper bound on detection rate based on item-representations; $\hat{k}^*$= upper bound on number of individual item representations VWM can hold (Eq. 4); **pure-item model:** $\hat{k}_p$= number of individual item representations VWM can hold (Rouder et al., 2011; Pashler, 1988). Given the different types of responses, different model parameters are extractable from the four experiments.

Table B1

Best Fitting Parameter Estimates

Estimate	Pure-Item	Item+Ensemble	Swapping
$\hat{k}$	2.78	2.39	2.92
$\hat{g}$	.10	.10	.00
$\hat{s}_e$	-	1.01	-
$\widehat{sw}$	-	-	.10

Note. $\hat{k}$ = number of individual items in VWM; $\hat{g}$ = guess rate; $\hat{s}_e$ = average size of ensemble representations; $\widehat{sw}$ = swapping rate. See text for model details and interpretations.

1

Figure Captions

Figure 1. Examples of 'change' trials that do (A – C) or do not (D) provide opportunities for forming ensemble representations based on feature/category repetitions. (A) Displays reproduced from the single-probe change-detection task in Brady and Alvarez (2015b, Fig. 2; adapted with permission © 2014 American Psychological Association): the four cubes on the left and the four characters on the right may be grouped, so that a change can be detected based on the fact that a character appears where there was a (cluster of) cube(s) in the memory display. (B) Displays from the whole-display change-detection task employed by Brady and Tenenbaum (2013, Fig. 3; adapted with permission © 2012 American Psychological Association): based on feature homogeneity and spatial proximity, the objects cluster into separable groups. (C) Standard colour-patch change-detection with colour repetitions reproduced from Brady and Tenenbaum (2013, Fig. 13; adapted with permission © 2012 American Psychological Association): the three violet squares on the right might be represented as one ensemble, so that the absence of a three-square violet group in the test display would indicate that a change has occurred; (D) Standard colour-patch change detection without colour repetitions as employed in the present study.

Figure 2. Example of a change trial in the present study. In all experiments, participants had to memorize the 6 colours from the memory array. In Experiments 1, 2, and 3a, one object changed from memory to test on half the trials, and participants indicated whether a change had occurred (Exp. 1 and 3a) or they mouse-clicked the changed object or clicked the fixation cross in case they had not perceived a change (Exp. 2). In Experiment 1, participants additionally selected the changed object (if there was a change; as illustrated by the figure panels in square brackets) after having received feedback on their response accuracy on the detection part of the task. In Experiment 3b, a change occurred on all trials and participants had to simply select the changed object without any decision regarding the presence of a change. Note that even if they had not perceived the change (and had responded "no change" in Exp. 1), participants were still required to indicate the location of the change in Experiment 1 and in Experiment 3b.

Figure 3. Spatial bias of erroneous selections. Whenever participants did not chose the correct location of the change, they had a slight tendency to select an object that was close to the change (closest object: ordinal position 1) rather than an object that was far from the change (most distant object: ordinal position 5), at least on trials where they detected a change (hit trials). The dotted line at 0.2 indicates the chance performance that would be expected if participants had no information on the location of the change. Error bars are 95% within-subject confidence intervals for the main effect of ordinal position (Jarmasz & Hollands, 2009; Loftus & Masson, 1994).

Figure 1

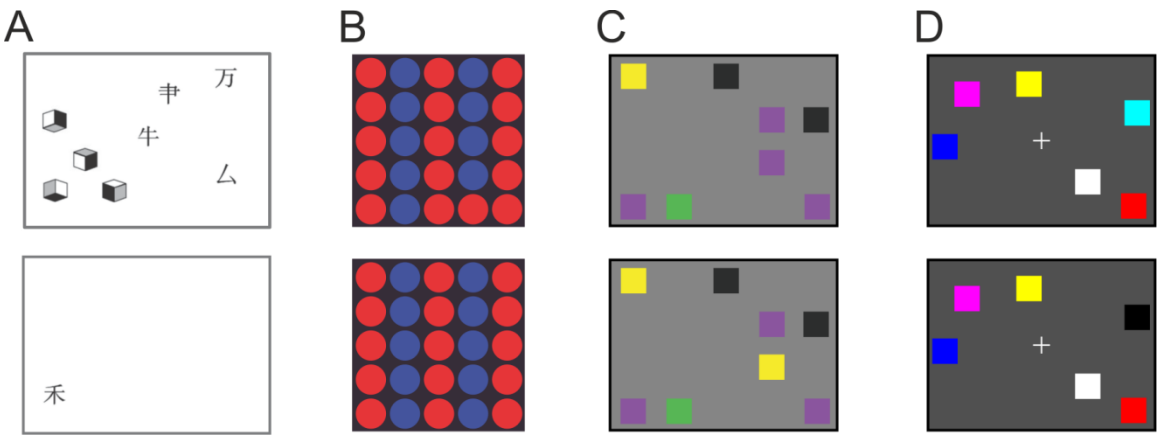


Figure 2

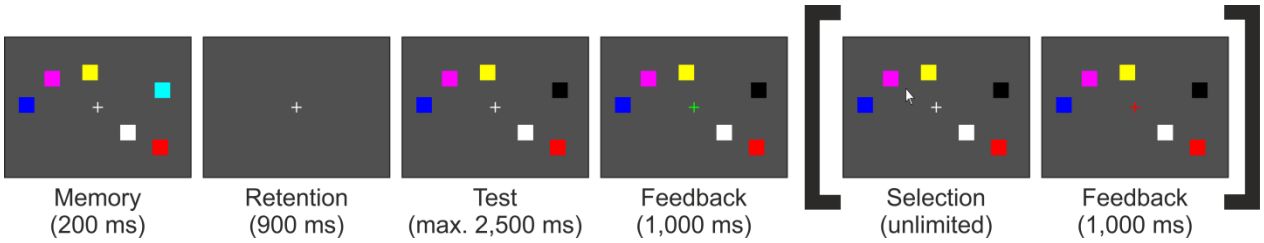
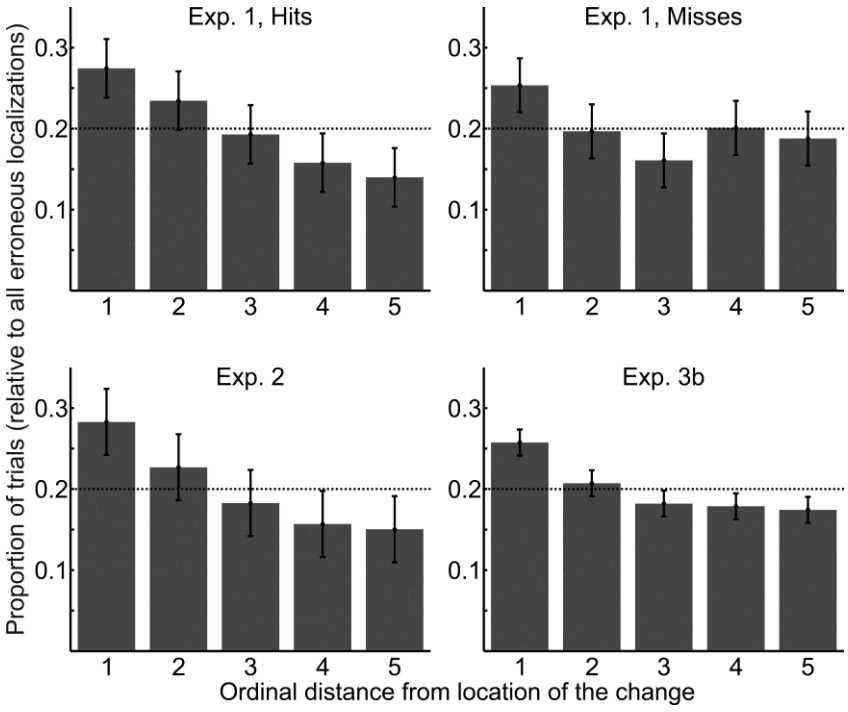

Memory
(200 ms)
Retention
(900 ms)
Test
(max. 2,500 ms)
Feedback
(1,000 ms)
Selection
(unlimited)
Feedback
(1,000 ms)

Figure 3


Exp. 1, Hits
Exp. 1, Misses
Exp. 2
Exp. 3b
Proportion of trials (relative to all erroneous localizations)
0.3
0.2
0.1
1
2
3
4
5
Ordinal distance from location of the change